Detecting environmental change: How many samples are required?

M. J. Gardner[1] and S. D. W. Comber[2]

[1] Atkins Limited, 500, Park Avenue, Aztec West, Almondsbury, Bristol BS32 4RZ, UK. Tel: +44(0)7834 506 966 Fax: +44 1454 663333; email: michael.gardner@atkinsglobal.com

[2] Plymouth University, B525 Portland Square, Drake Circus, Plymouth, Devon, PL4 8AA UK

Abstract

One of the most important functions of environmental monitoring is the detection of change. This can be the delineation of deteriorating circumstances or the identification of the success of remedial measures. The design of effective monitoring of change (and hence the optimisation of resources devoted to monitoring) relies on appropriate replication – knowing how many samples are required. Lack of information on the variance of the measured parameter is often a barrier to determining the optimum sampling strategy. An important new information resource on within-site variance of the concentrations of over 60 trace substances in wastewater treatment works effluents has been provided by the UK water industry research programme. This paper makes use of this resource in order to explore the potential to design monitoring programmes that will be capable of demonstrating the success of planned remedial measures that will be implemented in the coming years. Two approaches to experimental design (simple before-and-after sampling and detection of trends via correlation) are examined. It is concluded that for programmes involving numbers of samples of less than 30 the detection of a change in concentration of less than 50% might be very challenging for many of the trace substance of greatest interest. Knowledge of the difficulty of the task in hand should make it possible to design programmes that optimise the use of resources and the approaches taken, such that effects of interest are detected as soon and as economically as possible.

Keywords: detecting change, sewage effluent quality, trace chemicals

Introduction

The design principles for environmental monitoring programmes are well understood and widely documented. They include precepts such as: determining clear objectives, spending as much time as necessary to define the questions to be answered, setting data quality and quality control requirements, never beginning without knowing how the data will be analysed. Unfortunately, these and many other key principles are more often ignored than adhered to by programme designers. Often, the tendency is to proceed on the basis of what has been done before or to be guided primarily by what can apparently be afforded.

A. J. Underwood[1] has noted that "Much sampling to detect and quantify human environmental disturbances is flawed by a lack of appropriate replication". Whilst Underwood refers to replication in the spatial sense, similar considerations apply in the temporal realm, where trends are of interest. The point is that the determination of the number of samples to be taken or measurements made as a key design decision. Taking too many samples wastes resources, but, more commonly and more seriously, taking too few leads to lack of clear conclusions and in many cases the waste of even more time and effort. In fairness to programme designers, it is rarely possible for particularly well-informed decisions to be taken concerning optimising numbers of samples, because information about the variance of environmental parameters is unavailable or unreliable.

The UKWIR (UK Water Industry Research) chemicals investigation programme (CIP) [2,3] is one of the largest environmental monitoring exercises undertaken in the UK over the past 10 years. As such, it has developed through several phases. Initially, (2008-2013) the aim was to prioritise the risks posed by a range of recently regulated trace contaminants. Once the substances of greatest importance had been shown[2] to include fire retardants (fluorinated and brominated), tributyltin, polynuclear aromatic hydrocarbons, cypermethrin and steroids the next step in the Programme (2014 -2020) was to determine the likely numbers of sites at which there was the likelihood of failure to comply with environmental quality standards (EQSs). This current programme covers over 40 trace contaminants and is now reporting approaching a million results for over 170 wastewater treatments works (WwTW) effluents and surface waters. The next stage, having established that some potential problems can be categorised as widespread, whilst others are more localised, will be to apply control measures and to undertake monitoring to demonstrate how effective these have been. In short, the future requirements of the CIP include monitoring of change or trends, in both effluents and in surface waters. The data already accumulated by the CIP represents an invaluable resource from which to plan future monitoring of environmental change.

As noted above, much has already been written on the topic of the design of monitoring programmes – notably the publication by Ward et *al*[4]. Rather than repeating the sound advice already provided, this paper is an exploration of how the information resource provided by the CIP might be used and what it might imply with respect to effective monitoring of future measures to control pollution involving the different substances of interest.

Detection of a change

The simplest to attempt to detect change is to undertake two sets of analyses, one to establish initial conditions and a second set to determine whether or not any later change is detectable as statistically significant at an appropriate level of confidence.

An approach that concentrates all the monitoring effort into only two occasions and analysing approximately the same number of samples on each occasion is the simplest strategy. It does have the weaknesses that no information on the nature of any actual or any potential future trend is obtained as this is based on only two sampling occasions, and that it is not possible to know in advance when the second set of analyses should be carried out. This might mean that that analysing too soon fails to achieve detection or that putting off analysis delays the detection of an important or much needed change. The choice of sampling times also might be affected by shorter-term, non-permanent (e.g. seasonal) changes that should not be allowed to confuse the issue.

Nevertheless, having decided to adopt this simple approach, the next step is to consider the capability – the power - to detect changes. This is characterised by;

a) The size of change that occurs – or is of interest. The "effect";

b) The variance of the quantity being determined – usually expressed as the coefficient of variation (CoV, the ratio of standard deviation/mean). This is the area in which CIP data can provide an assessment of hitherto unrivalled accuracy. For trace metals, such as copper and zinc and sanitary parameters such as BOD, CoV values in the range 0.5-0.7 are typical. For trace organic substance higher CoV values in the range up to 1 or higher are not uncommon. Table 1 lists the within-site CoV values in WwTW effluents for substances of interest in the CIP. It should be noted that between-site CoV (also characterised in CIP) is a statistic that might be of interest in distinguishing between sites, but that it has no application here;

c) The required power – this is the probability of correctly detecting a true effect as statistically significant. Power is influenced by the choice of significance level for the test, the size of the effect being measured, and the number of measurements involved. Power is an expression of how sure we wish to be that the effect of interest will be detected; it is determined by the specified number of samples. A power of 0.8 is often the starting point of a statistical design[5]. This means that an actual effect of interest will be detected (in the long run) in 4 out of every five tests carried out to the design adopted;

d) The chosen level of statistical significance at which detection is recognised.

The statistical power of a test to detect an effect is the probability that the null hypothesis (i.e. that there is no trend) will be rejected when there is, in fact, an effect. It is the probability of making the correct decision. Power, and the number of samples required and hence the cost of sampling and analysis are inversely related, as is illustrated below.

Fig. 1 shows the relationship, for a simple two occasion test, between numbers of samples required and CoV – based on a required power based on a 0.8 and detection at a level of significance of p=0.05. Three sizes of change[1] are illustrated - reductions from a starting value of 1 to respectively 0.75, 0.5 and 0.25.

The statistical basis for the estimation of the power of a "t" test, based on the non-central t distribution, has been described by Cohen[5] and Chow *et al*[6]. This approach is a variant of the familiar "Student's t" test for the difference between two mean values. The Student's t distribution characterizes how the t test statistic is distributed when the null hypothesis is assumed to be true, i.e. that there is no difference. The non-central t distribution is a generalised version of the t distribution that shows how the t test statistic is distributed when the alternative hypothesis is assumed to be true. As such it is useful in calculating the power of the t tests and in estimating the numbers of samples required to detect a specified change at chosen levels of power and confidence. The assumptions and limitations of the approach are those associated with "t" tests – sufficient numbers of samples used to estimate mean values should ensure the assumption of approximate Normality is valid, but the simplest variants of the test rely on the data being uncorrelated and that the sample sizes are approximately equal. Essentially, the critical assumption is that surrounding the accuracy of estimates of CoV values which underlines the use, described below, of the large CIP data set in this context.

The calculations of sample numbers shown in Fig. 1 are based on this approach and on guidance provided in the Real Statistics Resource Pack (2013-2015)[7].

Fig. 1 illustrates the infeasibility of the detection of small changes. For a reduction of only a quarter, the number of samples increases sharply (to levels that are practically unrealistic?) as CoV rises to values greater than 0.4. Detection of a reduction to ½ of the starting value (an effect of 0.5) is evidently more achievable within the resources available to many programmes, though up to 40 samples might be required (40 samples analysed at the beginning of change assessment and 40 after) for some of the determinands with higher CoV. The curve for a reduction of 0.75 offers some prospect of ready detection, but then a fall as large as this might not be easy to achieve by the reduction measures used, unless the period of reduction is unduly long. This would imply that such measures would need to be applied for a long time before it could be shown that they had been successful.

Detection of changes for CIP determinands

A description of the CIP sampling regime is required to provide background to the use of CIP data to estimate the numbers of sample that might be required to detect changes in contaminant concentrations in sewage effluents resulting from future control measures. In this context sampling at each wastewater treatment works was carried out over a period of two years between 2015 and 2017, with effluents being sampled for over 40 determinands. Sampling was on a stratified random basis with site visits being made at approximately 2-monthly intervals throughout the period. Single samples were taken. At least 20 samples were collected at each site. The variance of results for each substance was then calculated and summarised as a CoV value that comprised a number of sub-components – variance of analysis, of the sampling process and the true variance of

[1] labeled as "reduction" but increases would be equally valid

effluent quality over the two-year period. Separation of these elements was not possible, though it is reasonable to assume from previous analyses[2] and quality control data that the variance of effluent quality predominated.

The CoV values between sites were than examined and the 25th, 50th and 75th percentiles noted for each determinand.

Fig. 2 (a and b) shows the numbers of samples required to detect, with a power of 0.8 at p=0.05, a 50% reduction in effluent concentration of the substances listed in Table 1. The upper panel (2a) shows sample numbers for substances with CoV values ranked in increasing order from calcium (CoV, 0.1) to BOD (CoV, 0.5). The lower panel illustrates required sample number for substance with median CoV values from 0.6 (reactive aluminium) to 1.7 (ibuprofen). All CoV values were estimated from CIP data based on at least 20 determinations from each of 170 wastewater treatment works. Estimated numbers of samples indicated by the columns are based on the median within-site CoV values from Table1 with lower and upper error bars respectively showing numbers required for CoV values at the 25%ile and 75%ile of within-site CoV values of the 170 works.

This illustration shows that statistical detection of change between the two sampling periods for sites of relatively low within-site CoV (for example, for a CoV of less than 0.6) might be achieved for many substances by the analysis of 20 to 30 samples on a simple before and after basis. Fig. 2 also illustrates the high sensitivity of the required number of samples to even a small increase of CoV above this nominal threshold. Conversely, increases in CoV markedly increase risk of failing to detect changes for a given number of samples.

It might be judged that the majority of substances and sites listed on the x-axis of Fig2(a) might fall into this low CoV category. However, it should also be noted that a quarter of sites will be subject to CoV values of greater than the 75%ile CoV, so there is not complete confidence of detection of the 50% change at all sites even for these most favourable cases.

Consideration of the higher CoV substances listed in Fig. 2(b) leads to the conclusion that even where the CoV is not greater than the median (i.e. half of sites) detection of change will in many cases require considerably more than 20 samples. For the substances in the right-hand half of Fig. 2(b) detection will be reliant on the CoV being lower that that commonly encountered. Given these less than encouraging predictions it might be worth considering the potential performance of alternative approaches. One such, trend detection by correlation, is discussed below.

Trend detection - use of rank correlation

Calculation of correlation between time and concentration is a well-established and robust method of trend detection [8,9,10]. A non-parametric approach based on data ranking should generally be used in order that the magnitude of results (or the presence of "outliers") does not play a part in the assessment. The variant illustrated as an example below employs Spearman's rank correlation, though other approaches that are essentially equivalent in outcome, if not in methodology, might be used. These include the calculation of Kendall's tau statistic[10]. The advantage of rank correlation over correlation of the untransformed data (Pearson r) is that there

is no implied assumption of linearity. The approach can, however, became complicated if there are tied ranks, but this is not necessarily a problem with all data sets. In order to obtain the Spearman's rho coefficient, the concentration and time values are assigned a rank value and a conventional (Pearson) correlation calculated on the ranks. This approach This can then be assessed for statistical significance. The resulting tests show the statistical significance of any monotonic trend – i.e. one that involves both time and concentration changing concurrently, but not necessarily at the same rate (as they would in a linear trend).

The key question to be asked in trend detection is slightly different from that for the detection of a change. It is more along the lines of "for a given real trend and monitoring frequency, how long would it take for a significant change to be detected?" An example can illustrate the process. Suppose the underlying trend is one in which concentration exponentially decreases with a half-life of 5 years and samples are taken and analysed once every two months (illustrated for CoV of 0.6 in Fig. 3). Assessments of the significance of correlation are then made each year, as the data series extends, until a statistically significant correlation is evident. Assessment need to be made on an annual basis in order to negate seasonal effects.

The question of power can be answered by Monte Carlo simulation. By generating a thousand simulated series of bi-monthly samples for an extended period (in this case for 11 years, but the length of the series does not matter provided it is sufficiently long). It can be assessed, in the long run, how long it takes for significance to be achieved. The fact that the correlation is carried out at each year end, incorporating all the available data (unlike the simple differences between successive tests on average values) offers the prospect that the cumulative effect of change might provide an increased power that the simpler tests cannot achieve. Figure 3 illustrates a sample set of data fitted with a LOESS smoother curve. The LOESS smoother is a means of producing a visualisation of trend data in which the position of the LOESS curve at any point is determined by a weighted regression based on nearby points – the weighting decreasing with distance from the point itself [11, 12]. The simulation results showing year of detection of a statistically significant trend and numbers of sample analysed for different data CoV values are shown in Table 2.

This rank correlation method has the advantages over simple difference methods in that it is not dependent on having a precise estimate of starting conditions, it provides the opportunity for convincing non-statistical visualisations of data to illustrate change (of course, supplemented by statistics) and, importantly, the approach to sampling provides a continuous analytical load that is more likely to lead to better analytical support and the opportunity to generate credible quality control data than might be the case for analyses carried out in discrete batches. The example above suggests that it might be unlikely to achieve reliable detection of trends resulting in a change of less than 50% for substances with CoV values in the 0.5-0.7– an outcome in broad agreement with the difference test. However, comparing the illustration in Fig. 1 for the detection of difference with the correlation data from Table2 shows the potential advantage of the correlation approach. For an actual change of 50% and a CoV of 0.8, 54 samples would be required (27 at the start of the test and 27 at the end) for detection of a difference, whereas the correlation method requires a series of 36 samples. Obviously, these are only illustrations, but they do suggest that a trend-based approach might perform better than difference methods for determinands of higher variance, owing to its accumulating power as it is continued.

Conclusions

It has been shown that detection of changes of less than 50% in the concentration of many trace contaminants in surface waters is unlikely to be achieved unless numbers of samples used in direct comparison studies are in the range of approximately 10 to 30 and that within-site CoV values are generally lower than 0.8. Sample numbers in the lower part (10-15) of this range might be appropriate for dissolved metals, PFOA, DEHP and several pharmaceuticals. Trace organic contaminants of current concern in the UKWIR CIP investigations, including PAHs, tributyltin, hexabromocyclododecane, steroids oestrogens and cypermethrin might require sampling rates approaching or exceeding 30 samples.

The results reported here suggest that the use of correlation based approaches to trend detection might be more powerful than simpler attempts to detect differences at intervals. There is also an implication that, in any future investigations of trends, merely deciding on a set number of samples to be collected and applying this to all substances might result in relatively early detection of change for some substances, but less success for others. A more rational approach might be to use the existing information on the variability of concentrations to determine the numbers of samples requires for low and high CoV determinands. Indeed, the variability of some substances might be such that trend detection could be ruled out as a practicable proposition before resources are wasted in a venture inconsistent with available resources. It might be that these findings are unlikely to influence the often-optimistic expectations of programme sponsors, however, they might be of value to monitoring programme designers in defining the scale of the task they might be required to address.

Acknowledgements - The author wishes to thank the co-ordinator of the CIP programme – UK Water Industry Research (UKWIR) for authorising the use of the information reported here, and the UK Water Utility companies Anglian, Dwr Cymru, Northumbrian, Severn Trent, Southern, South West, Thames, United Utilities, Wessex and Yorkshire Water for their efforts in generating it.

References

1. Underwood A.J.. (1994) On Beyond BACI: Sampling designs that might reliably detect environmental disturbances: Ecological Applications, 4, 3–15.

2. Gardner, M.J., Comber, S.D.W., Scrimshaw, M.D., Cartmell, E., Lester, J, and Ellor, B. (2012). The Significance of Hazardous Chemicals in Wastewater Treatment Works Effluents. Science of the Total Environment, 437, 363-372.

3. Hayward, K, (2016) Early insights from the UK's groundbreaking sewage assessment, Aqua Strategy October, accessed 08/06/2017 https://www.aquastrategy.com/article/early-insights-uks-groundbreaking-sewage-assessment

4. Ward, R. C., Loftis, J.C. and McBride G.B. Design of Water Quality Monitoring Systems. John Wiley & Sons, 1990 - Technology & Engineering – ISBN 0471283886, 9780471283881, 256 pp.

5. Cohen, J. (1988), Statistical Power Analysis for the Behavioral Sciences, 2nd Ed., New York: Lawrence Erlbaum Associates.

6. Chow, S., Shao, J. and Wang, H., (2002) A note on sample size calculation for mean comparisons based on noncentral t-statistics Journal of Biopharmaceutical Statistics, Vol. 12, No. 4, pp. 441–456.

7. The Real Statistics Resource Pack software (Release 4.3). Copyright (2013 – 2015) Charles Zaiontz. www.real-statistics.com.

8. Gauthier, T. D. (2001) Detecting Trends Using Spearman's Rank Correlation Coefficient. Environmental Forensics Volume 2, Issue 4, December 359-362.

9. Yue, S. Pilon, P. and Cavadias, G. Journal of hydrology (2002) Power of the Mann-Kendall and Spearman's rho test for the detection of trends monotonic trends in hydrological series. 259 254-271.

10. Helsel, D.R. & Hirsch, R.M. (1992). Statistical Methods in Water Resources, Elsevier, ISBN 0-444-81463-9, Amsterdam.

11. Cleveland, W.S. and Devlin, S.J. (1988) Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting, Journal of the American Statistical Association 83, 569-610.

12. NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, accessed 12/12/2017.

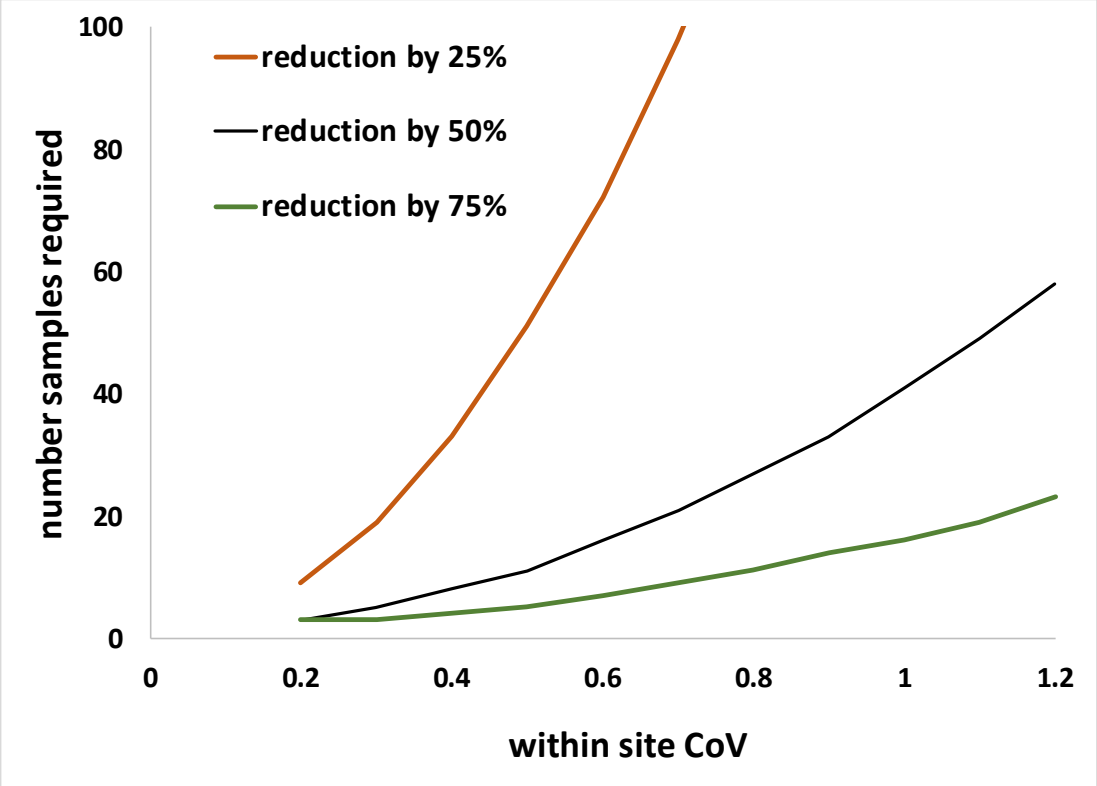
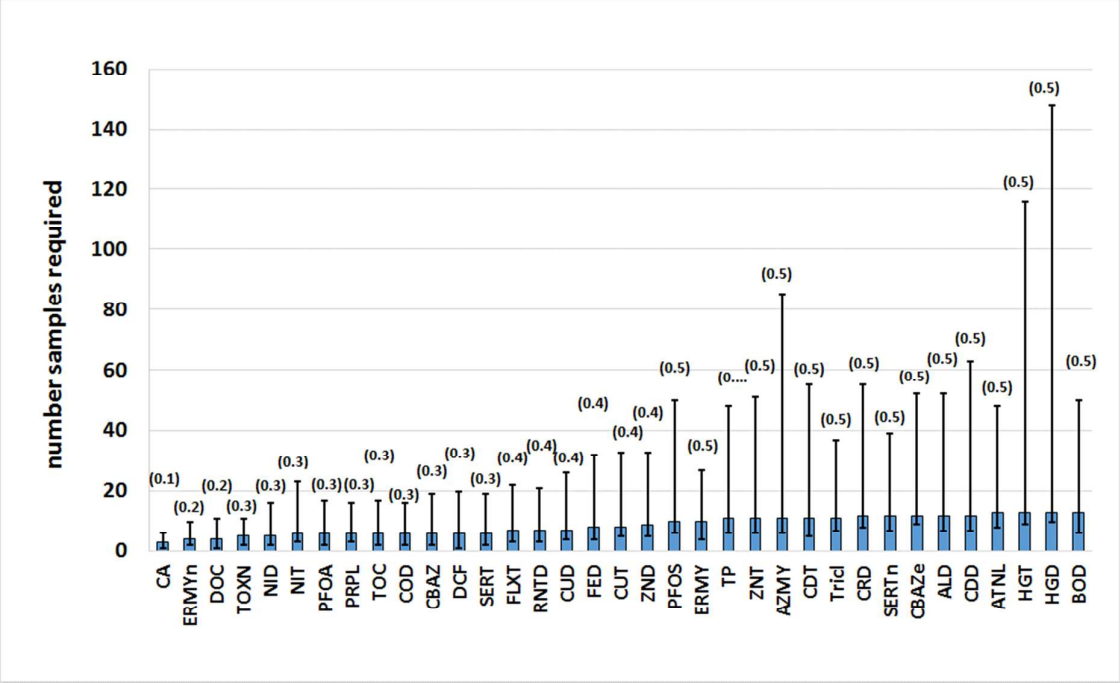
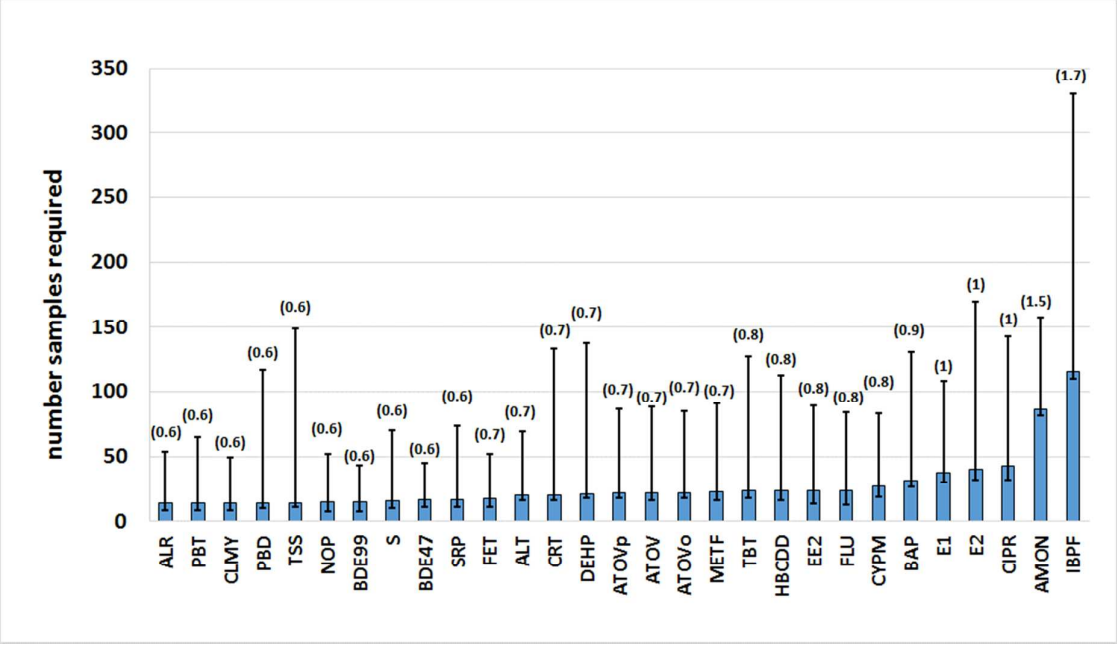


Fig. 1 Numbers of samples requires to detect by difference with power of 0.8 at a significance level of p=0.05



2a – median within-site CoV values from 0.1 to 0.6



2b - median within-site CoV values 0.6 to 1.7

Fig. 2 Estimated numbers of samples required to detect a 50% change in effluent concentrations

Estimated numbers of samples indicated by the columns are based on the median within-site CoV values from Table1 with lower and upper error bars respectively showing numbers required for CoV values at the 25%ile and 75%ile of within-site CoV values of the 170 CIP works. CoV median values are shown in brackets.

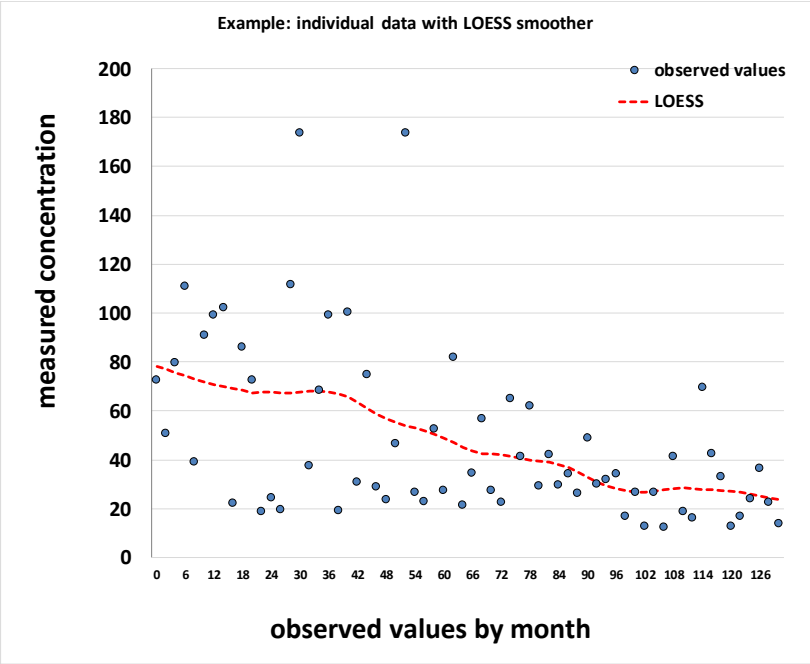


Fig. 3 Illustration of an exponentially decreasing trend (t½ =5 years) with variance corresponding to a CoV of 0.6

Table 1 Within-site CoV values for WwTW effluents

Substance	Code		With-site CoV values for effluents		
		Median Concentration units (µg/l) unless stated under "substance"	25%ile	50%ile	75%ile
nickel (dissolved)	NID	3.4	0.18	0.28	0.49
nickel (total)	NIT	3.8	0.20	0.32	0.63
lead (dissolved)	PBD	0.25	0.31	0.57	1.19
lead (total)	PBT	0.62	0.32	0.56	1.16
copper (dissolved)	CUD	4.3	0.19	0.37	0.67
copper (total)	CUT	7.6	0.20	0.40	0.77
zinc (dissolved)	ZND	23	0.26	0.43	0.75
zinc (total)	ZNT	32	0.27	0.49	0.99
cadmium (dissolved)	CDD	0.019	0.29	0.52	1.12
cadmium (total)	CDT	0.027	0.34	0.50	1.04
mercury (dissolved)	HGD	0.0020	0.20	0.53	1.84
mercury (total)	HGT	0.0040	0.22	0.53	1.61
iron (dissolved)	FED	87	0.21	0.39	0.75
iron (total)	FET	271	0.21	0.66	1.26
aluminium (dissolved)	ALD	14	0.29	0.52	0.99
aluminium (total)	ALT	45	0.36	0.68	1.31
aluminium (reactive)	ALR	7.6	0.38	0.56	0.82
chromium (dissolved)	CRD	0.28	0.23	0.50	1.03
chromium (total)	CRT	0.51	0.33	0.69	1.61
diethylhexylphthalate	DEHP	0.45	0.41	0.71	1.49
BDE 47	BDE47	0.00039	0.26	0.62	1.28
BDE 99	BDE99	0.00036	0.25	0.60	1.69
PFOS	PFOS	0.0053	0.22	0.46	0.99
PFOA	PFOA	0.0055	0.23	0.32	0.49
HCBDD	HBCDD	0.0074	0.41	0.76	1.81
nonylphenol	NOP	0.1047	0.22	0.58	1.10
tributyltin	TBT	0.00016	0.39	0.75	1.32
fluoranthene	FLU	0.0101	0.31	0.76	1.34
benzo(a)pyrene	BAP	0.0035	0.37	0.88	1.68
triclosan	Tricl	0.067	0.21	0.50	0.80
cypermethrin	CYPM	0.00014	0.35	0.80	2.33
total suspended solids	TSS mg/l	8.4	0.39	0.57	0.91
ammoniacal nitrogen (as N)	AMON mg/l	0.52	0.64	1.46	2.68
total oxidised nitrogen (as N)	TOXN mg/l	20	0.20	0.28	0.34
Biochemical Oxygen Demand	BOD mg/l	3.5	0.37	0.54	0.95

Chemical Oxygen Demand	COD mg/l	31	0.23	0.32	0.47
total phosphorus (as P)	TP mg/l	1.45	0.28	0.49	0.95
soluble reactive phosphate (as P)	SRP mg/l	1.02	0.32	0.62	1.28
total organic carbon	TOC mg/l	11.3	0.22	0.32	0.50
dissolved organic carbon	DOC mg/l	8.8	0.14	0.24	0.37
calcium	CA mg/l	81.4	0.08	0.11	0.16
sulphide	S mg/l	0.0065	0.17	0.61	1.71
oestrone	E1	0.0043	0.51	0.958	1.72
17β oestradiol	E2	0.0007	0.34	0.995	2.27
17α ethinyloestradiol	EE2	0.00020	0.49	0.759	1.58
diclofenac	DCF	0.28	0.28	0.34	0.56
ibuprofen	IBPF	0.11	0.18	1.71	4.39
atorvastatin	ATOV	0.10	0.48	0.73	1.23
ortho-hydroxyatorvastatin	ATOV o	0.18	0.42	0.73	1.18
para-hydroxyatorvastatin	ATOV p	0.21	0.47	0.72	1.28
propanolol	PRPL	0.17	0.19	0.32	0.46
atenolol	ATNL	0.32	0.28	0.53	0.92
erythromycin	ERMY	0.35	0.33	0.46	0.63
norerythromycin	ERMY n	0.05	0.00	0.21	0.33
azithromycin	AZMY	0.20	0.28	0.49	1.36
clarithromycin	CLMY	0.40	0.32	0.57	0.83
ciprofloxacin	CIPR	0.14	0.21	1.03	1.93
metformin	METF	4.81	0.27	0.74	1.58
ranitidine	RNTD	0.55	0.23	0.37	0.56
carbamazepine	CBAZ	0.64	0.24	0.34	0.55
10,11-epoxycarbamazepine	CBAZe	0.12	0.17	0.52	0.98
sertraline	SERT	0.06	0.24	0.35	0.54
norsertraline	SERTn	0.03	0.29	0.51	0.80
fluoxetine	FLXT	0.05	0.24	0.37	0.58
benzotriazole	BZT	1.44	0.23	0.34	1.06
tolyltriazole	TZT	1.28	0.21	0.34	0.64

Table 2. Detection of trend by rank correlation testing

Data within-site CoV	0.20	0.40	0.60	0.80	1.00	1.20
Year of detection for power of 0.8 at p=0.05	2	4	5	6	7	7
Percentage decrease occurring by year of detection	24	43	50	56	62	62
Numbers of sample analysed prior to detection	12	24	30	36	42	42